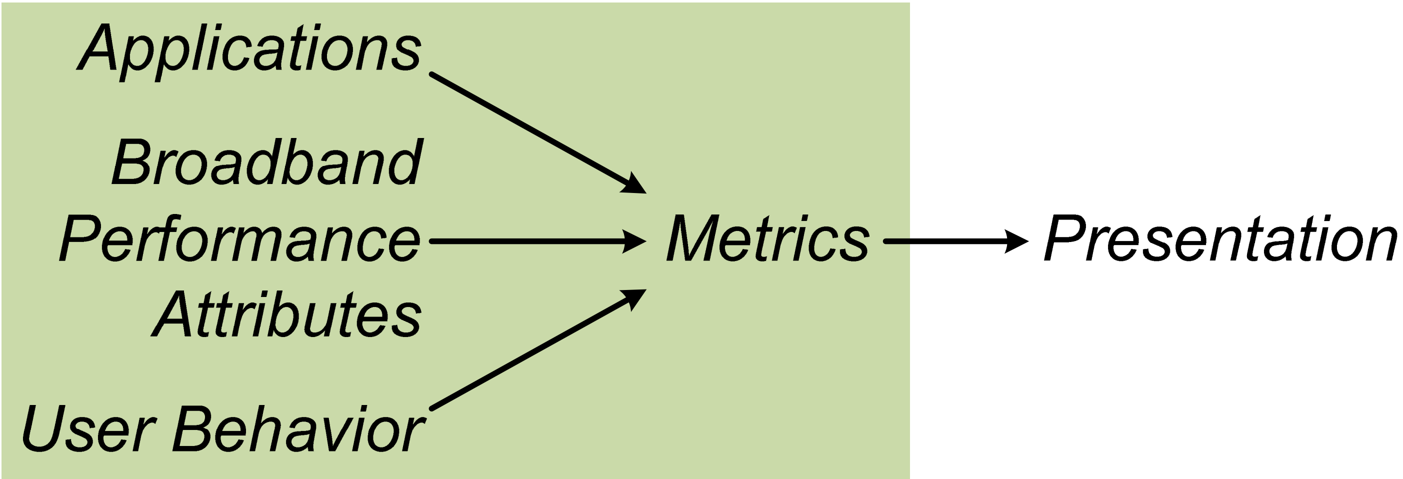
Application-Driven Broadband Metrics

January 2011

Goal of this Contribution

Identify the metrics on which to base Broadband Performance Results

Why Metrics?



- Presentation relies on metrics
- Metric choices determined by
  - Application needs
  - Broadband performance attributes
  - User behavior
- Define solid fundamentals and the presentation will follow

Application Classes

- Real Time (RT) Applications
  - VoIP (conversational voice); Video conferencing (conversational video)
- Near-RT Applications
  - Streaming video; Streaming audio
- Time Sensitive, Interactive
  - Gaming; Remote video (nanny cams, security)
- Transactions
  - Web browsing; E-commerce
- Background
  - Email; Peer-to-peer

Application/QoS Resources

Error tolerant	Conversational voice and video	Voice/video messaging	Streaming audio And video	Fax
Error intolerant	Command/control (e.g., Telnet, interactive games)	Transactions (e.g. E-commerce, WWW browsing, Email access)	Messaging, Downloads (e.g. FTP, still image)	Background (e.g. Usenet)
	Interactive (delay <<1 s)	**Responsive** (delay ~2 s)	**Timely** (delay ~10 s)	**Non-critical** (delay >>10 s)

Figure 2/G.1010 – Model for user-centric QoS categories

- Standards-based source documents
  - ITU Recommendation G.1010[1]
    - 2001 publication, some specifics outdated
  - 3GPP TS 22.105 V9.0.0[2]
  - Broadband Forum TR-126[3]
- Need to apply context
  - Requirements written for QoS-enabled networks cannot be applied directly to Internet, which is Best Effort
  - Applications designed for Internet tolerate wide variation in conditions

Broadband Performance Attributes

- Rate (subcategories and terms as used here)
  - Sustained rate – performance sustained over extended periods
  - Burst rate – initial performance exceeding sustained rate
  - Reliability – the probability with which a given rate is met or exceeded at a given point in time
- Latency
  - One way or round trip delay
- Jitter
  - Variation in delay
- Packet/frame loss
  - Packets or frames which are not received

User Behavior

- Concurrent applications
  - On the same device
  - On multiple devices in a subscriber's network
- Diurnal patterns
  - Time-of-day dependency

Application Classes and Performance Attributes

Real Time Applications

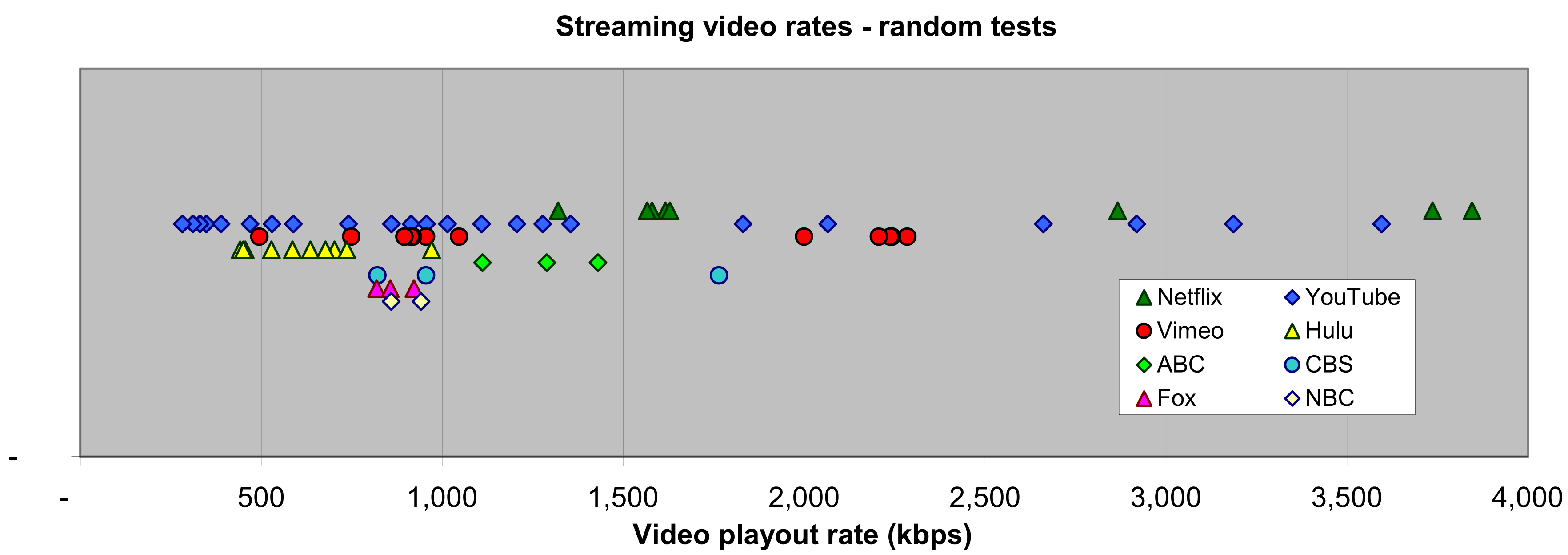
- VoIP and video conferencing applications
  - Continuous, near constant rate traffic
  - Network rate must be sustained and highly reliable
    - Performance below established rate for as little as 50 ms can cause dropouts
  - Latency should be low
    - 150 msec (one way, path from mouth to ear) is the threshold for natural interactive conversation
  - Jitter should be low
    - De-jitter buffers may be as short as 50 msec
    - Jitter exceeding buffer length causes buffer under- or over-runs, loss of data
- VoIP: low rate (≈100 kbps)
- Video conferencing: wide range of rates
  - Connect rate based on network speed
  - 256 kbps to 2 Mbps typical for consumer apps
  - HD teleconferencing (e.g., Cisco ūmi) requires at least 1.5 Mbps
  - Trend towards increasing rates

Near-RT Applications

- Streaming video
  - 51% of NA consumer Internet traffic in 2010 (and growing)[4]
  - Virtually all sent over TCP
  - Transfer rate dependent on client/server combination
    - Some send as fast as network (and TCP) allows
    - Some match playout rate after initial buffering period
  - Network rate must be sustained and reliable
    - Performance below playout rate for several seconds can cause freezes while re-buffering
  - Large file sizes
    - 10 minute video @ 1 Mbps → 75 Mbytes
    - 45 minute video @ 3 Mbps → 1 Gbyte
- Streaming audio similar (lower rates, smaller files)

Video Streaming Rates

- Downloads captured from popular sources
  - YouTube, Hulu, Vimeo, ABC, CBS, NBC, Fox, Netflix
  - Random, not exhaustive
- Playout rates from 280 kbps to over 3.8 Mbps
  - Wide variety across the range



Video Streaming Rates (2)

- Data collected from YouTube
  - Spring 2009:
    - Standard and “High Quality” rates from ≈200 to ≈700 kbps
  - Fall 2010:
    - 240p through 1080p rates from 280 kbps to 3.6 Mbps
    - 4k stream rate (not shown in chart) = 20.5 Mbps
- Rapid upward trend and growing range in playout rates
  - Fixed rate streaming test does not capture this range

Time-Sensitive Interactive Applications

- Gaming
  - Bursty traffic, low average rate (<<1 Mbps)
    - User inputs transmitted upstream
    - Multiplayer activity transmitted downstream
    - Game environment rendered locally
  - Latency is primary performance factor
    - Unless compensated by game server, player with lowest latency has advantage
    - High latency creates inconsistent experience between remotely located players
  - Jitter appears as variable latency
    - Inconsistent performance
  - Packet/frame loss causes retransmission, appears as variable latency
    - Inconsistent performance

Transactions

- Web browsing performance
  - Above about 5 Mbps, latency is much more important than rate
  - For an average web page
    - Doubling rate from 10 → 20 Mbps improves response time by ***less than ¼ second***
    - Decreasing round trip latency by 10 ms has the same effect on response time as increasing the rate by ***1 Gbps***
- Above conclusions are not intuitive!
  - Repeatedly justified in the literature[5,6,7]
  - See following slides

Transactions (2)

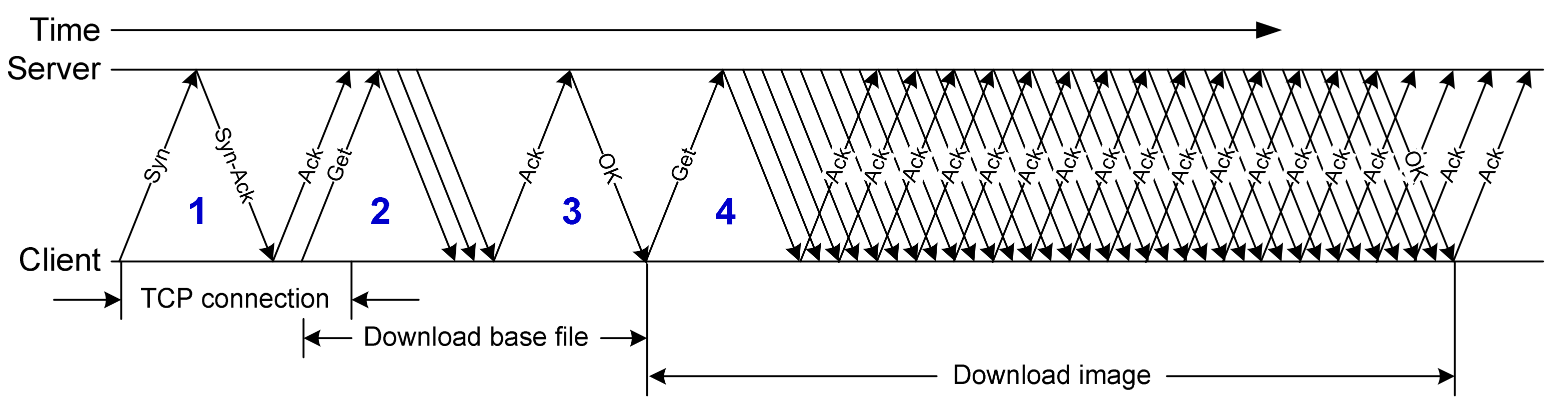
How long does it take to load a web page?[5,6]

$$R \approx \frac{Size}{Bandwidth} + Turns \cdot RTT + Cs + Cc$$

- Parameters
  - *R* = page load time
  - *Size* = total data to be transferred
  - *Bandwidth* = speed between client and server
  - *Turns* = effective # of round trips
  - *RTT* = Round Trip Time
  - *Cs* = server processing time
  - *Cc* = client processing time

Transactions (3)

- Effective turns
  - Effective number of round trips waiting for: DNS responses; establishing TCP connections; HTTP Gets; TCP slow starts; etc.
  - Less than total number of round trips due to: TCP windowing; parallel TCP connections; etc.
- Simplified example below requires 4 effective turns



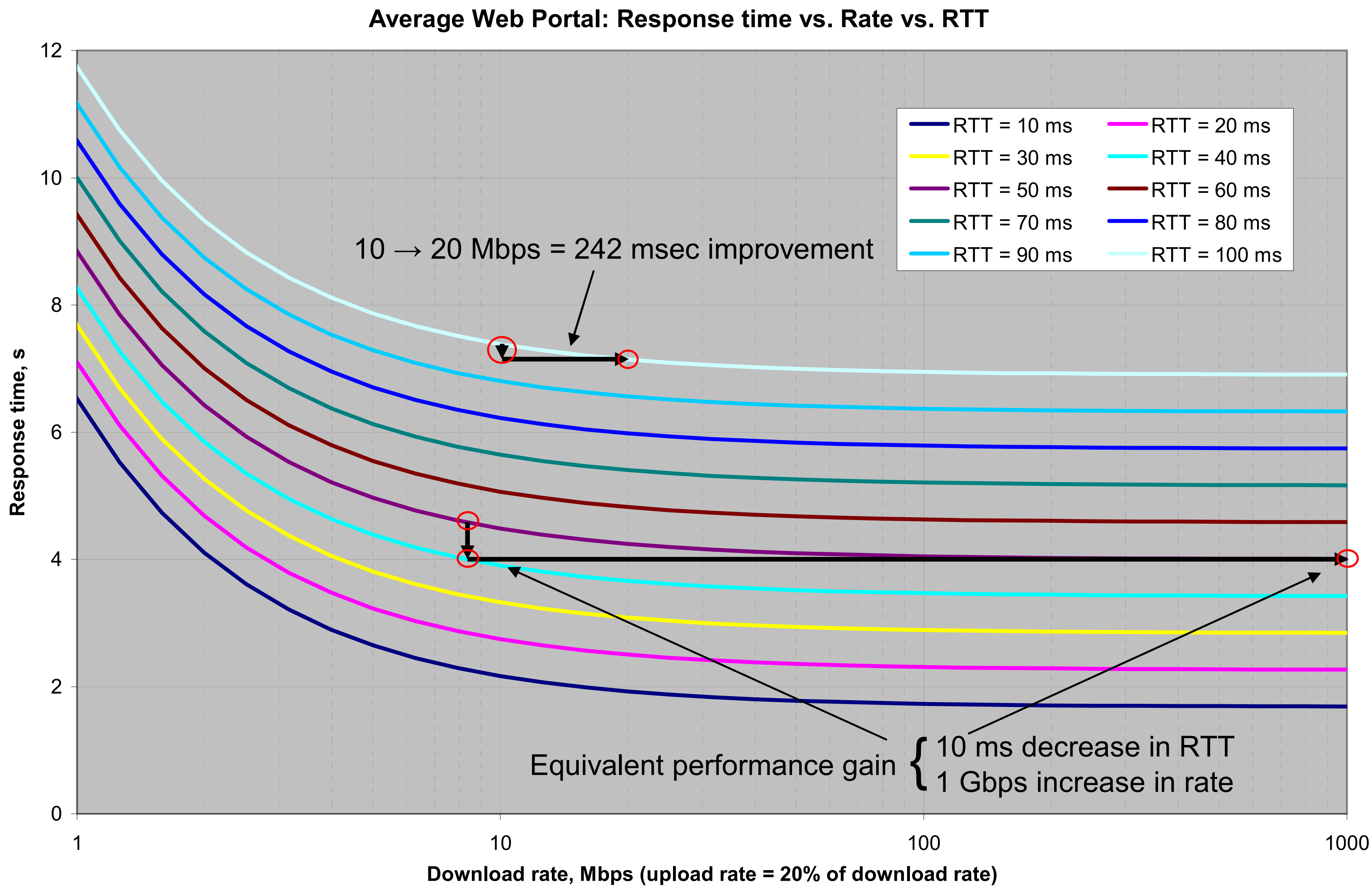
Transactions (4)

$$R \approx \frac{Size}{Bandwidth} + Turns \cdot RTT + Cs + Cc$$

- Two network dependencies for browsing performance:
  - Rate (*Size* / *Bandwidth*)
  - Latency (*Turns* x *RTT*)
- Latency is multiplied by the effective number of turns required to load a web page
  - This factor sets a floor on response time which cannot be bettered ***even with infinite rate***
- Web page statistics
  - 2001 survey[5] of Keynote Business-40 web sites
    - Average size ≈ 100 kBytes
    - Average effective turns ≈ 40
  - 2009 survey[8] of top 25 web sites per ranking.com (portal sites)
    - Average size ≈ 480 kBytes
    - Average effective turns ≈ 58

Transactions (5)

Effects of Rate and RTT on Web Page Response Time



Background Applications

- Email
  - Small file sizes: no significant rate dependency
  - Non-RT: tolerant of latency and jitter
  - TCP: tolerant of packet/frame loss
- Peer-to-Peer
  - Large file sizes: Higher sustained rate improves transfer times
  - Non-RT: tolerant of latency and jitter
  - TCP: tolerant of packet/frame loss

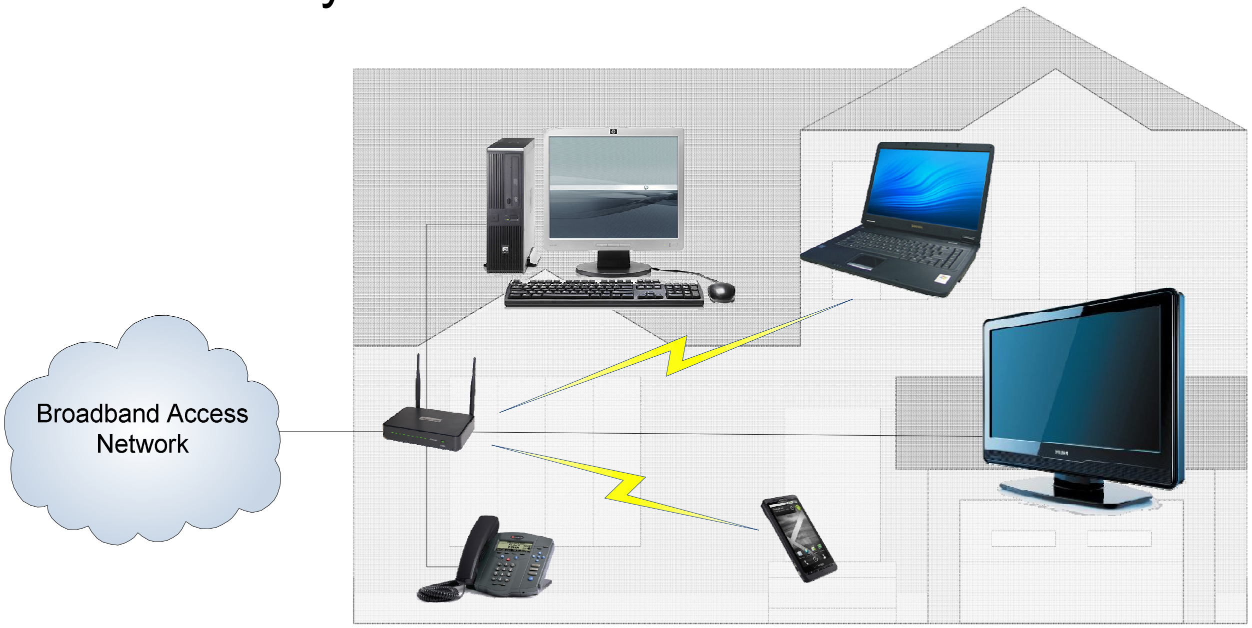
Application Performance Factors

Applications	Rate	Latency	Jitter	Packet/frame loss
VoIP, video conf.	• Video rates from 256 kbps to >2 Mbps and increasing – Sustained rate critical, must be very reliable – Burst rate: N/A	• Important • <150 msec	• Important • ≤ size of jitter buffer	• Tolerates moderate loss (≈1%)
Video (or audio) streaming	• Rates from 256 kbps to 4 Mbps and increasing – Sustained rate critical, must be reliable – Burst rate helps with initial buffering	• Tolerated	• Tolerated	• Tolerates moderate loss (≈1%)
Gaming	• Secondary to latency in importance • Traffic is bursty, average <<1 Mbps	• Important – As low as possible	• As low as possible	• As low as possible
Web browsing	• Rates (sustained or burst) above ≈5 Mbps have little effect on response time	• Can be more important than rate	• Tolerated	• Tolerated (TCP)
Email	• Rates (sustained or burst) have little effect on response time	• Tolerated	• Tolerated	• Tolerated (TCP)
Peer-to-peer	• Higher sustained rate improves transfer time • Burst rate: N/A	• Tolerated	• Tolerated	• Tolerated (TCP)

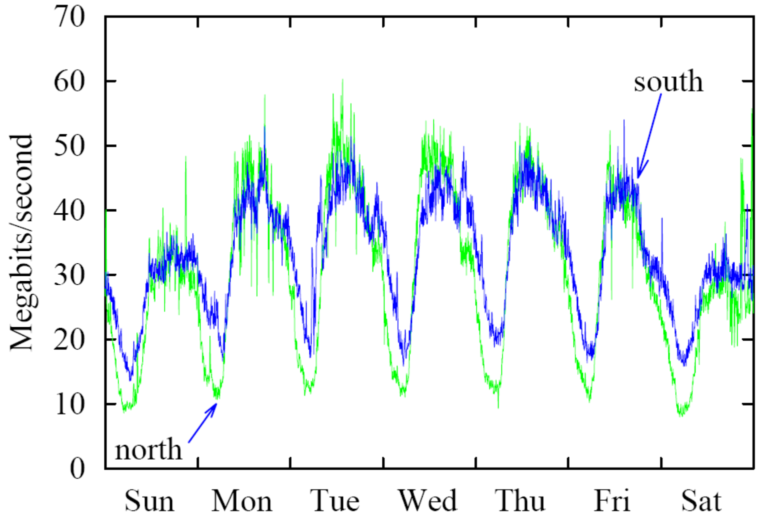
User Behavior and Performance Attributes

Concurrent Applications

- Devices support multiple applications in parallel
- Home networks support multiple devices in parallel
  - Multiple video streams (to PCs and TVs)
  - Web browsing
  - VoIP
- Higher rate performance becomes more important to support concurrent flows
  - Consumer education is important
- Concurrent flows may inhibit “burst rate” features



Time of Day Variation



- Diurnal traffic patterns are well known (above example from [9])
- Performance during busy hour may be significantly different from 24-hour average
  - Much heavier loading on network
  - Higher probability of congestion
- Separate results for peak and off-peak periods
  - 24-hour averages are less informative

Metrics

Metrics

- SamKnows tests are capturing data for all metrics listed below
- Sustained rate (downstream and upstream)
  - Sustained (30 second) metrics from download and upload speed tests
    - If first 5 seconds is significantly different (burst rate) may need to exclude it from sustained metric (use seconds 5-30)
  - ***Critical for: VoIP; video conferencing; streaming media; multiple concurrent apps***
  - Helpful for: gaming; web browsing; large file transfers
- Burst rate (downstream and upstream)
  - Burst (5 second) metrics from download and upload speed tests
  - Helpful for: streaming media; gaming; web browsing
- Latency
  - From UDP latency/loss tests
  - ***Important for: VoIP; video conferencing; gaming***
  - Helpful for: web browsing
- Jitter
  - From UDP latency/loss tests
  - ***Important for: VoIP; video conferencing; gaming***
- Packet/frame loss
  - From UDP latency/loss tests
  - Helpful for: VoIP; video conferencing; streaming media; gaming

Metrics vs. Presentation

- Metrics: inputs for analysis
  - Detailed
  - Comprehensive
  - Technical
  - Too complex for presentation "as is" to consumers
- Presentation: results of analysis → published data
  - Informative
  - Understandable
  - Usable
  - Simple
- Data values may help determine presentation
  - Example: latency, jitter, loss values may lie within acceptable ranges and may not need emphasis (e.g., Ofcom report[10])

Issues and Caveats

- Latency, jitter, packet/frame loss require cautious approach
  - Example: latency
    - Meaningful metric is end-to-end but distance (nodes, miles, etc.) between every pair of endpoints is different
    - Latency across access/aggregation network rarely represents end-to-end experience
    - May be best to assess against well-documented threshold (e.g., 150 msec for interactive conversation[1,2,3])
- Internet is Best Effort
  - Beware of wording that leads people to infer otherwise!
- Presentation should emphasize busy hour performance
- Terms (“reliable,” “highly reliable”) require definition

References

1. ITU Recommendation G.1010, “End-user multimedia QoS categories,” November 2001.
2. 3GPP TS 22.105 V9.0.0, “3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Service aspects; Services and service capabilities (Release 9),” December 2008.
3. Broadband Forum, Technical Report TR-126, “Triple-play Services Quality of Experience (QoE) Requirements,” December 13, 2006.
4. Cisco, “Cisco Visual Networking Index – Forecast and Methodology, 2009-2014,” June 2, 2010.
5. Sevcik, P., and Bartlett, J., “Understanding Web Performance,” NetForecast Report 5055, October 2001.
6. Savoia, A., “Web Page Response Time 101,” STQE Magazine, Vol. 3, Issue 4, July/August 2001, pp. 48-53.
7. Cheshire, S., "Latency and the Quest for Interactivity". *White paper commissioned by Volpe Welty Asset Management, L.L.C., for the Synchronous Person-to-Person Interactive Computing Environments Meeting, San Francisco*, November 1996.
8. ADTRAN, “Defining Broadband: Network Latency and Application Performance,” comments to FCC, June 2009.
9. Thompson, K., Miller, G., Wilder, R., “Wide Area Internet Traffic Patterns and Characteristics,” IEEE Network, November/December 1997.
10. Ofcom, “UK Broadband Speeds,” May 2010.